Trans Media Relevance Feedback for Image Autoannotation

Thomas Mensink, Jakob Verbeek, Gabriela Csurka

Trans Media Relevance Feedback for Image Autoannotation

Thomas Mensink[12]
thomas.mensink@xrce.xerox.com
Jakob Verbeek[2]
jakob.verbeek@inrialpes.fr
Gabriela Csurka[1]
gabriela.csurka@xrce.xerox.com

[1] TVPA - Xerox Research Centre Europe
Meylan, France

[2] LEAR - INRIA Rhône-Alpes
Montbonnot, France

Abstract

Automatic image annotation is an important tool for keyword-based image retrieval, providing a textual index for non-annotated images. Many image auto annotation methods are based on visual similarity between images to be annotated and images in a training corpus. The annotations of the most similar training images are transferred to the image to be annotated. In this paper we consider using also similarities among the training images, both visual and textual, to derive pseudo relevance models, as well as cross-media relevance models. We extend a recent state-of-the-art image annotation model to incorporate this information. On two widely used datasets (COREL and IAPR) we show experimentally that the pseudo-relevance models improve the annotation accuracy.

1 Introduction

In this paper we address the problem of image auto-annotation, where the goal is to predict relevant keywords from a finite vocabulary given a new image. These keyword predictions can then be used in tools for clustering, classification, retrieval and visualization. These tools are important to explore large quantities of images on photo sharing sites or in desktop photo management applications.

Image auto-annotation is closely related to image categorization, in the sense that both methods learn from existent labeled data the prediction of tags for unlabeled images. The difference resides mainly in the used dataset. In image categorization the set of labels is relatively small and predefined, while in image auto-annotation it is large and might even evolve in the case of dynamic databases (photo sharing, image repositories). Also, in image categorization the training set is often well structured, with generally all images completely annotated. On the other hand, auto-annotation is the process to extract potential labels for a new image, from a generally unstructured and often noisy data set.

One obvious solution to auto-annotation is hence to use image categorization techniques. For each (possible) keyword we can gather a set of positive (and negative) examples and train a keyword specific classifier. However this might be costly in case of large and dynamic image sets, and few of the actual systems scale well to large amount of classes. While recent techniques tend to address the problem of large scale image categorization, they either

consider mono-labeled data (as for the ImageNet dataset [13]) or a large dataset for only a relatively small number of classes as in [23].

An alternative solution is what is called tag propagation. The idea here is that considering the test images, similar images are gathered from the data set and the annotations are deduced from analyzing the tags, annotations or the text around these top retrieved images. Then one can either directly deduce the most relevant concepts/keywords to tag the test image [16, 28] or to learn a discriminative models in neighborhoods of test images [30]. Recently, these nearest neighbor type methods have shown excellent performance for auto-annotation [11, 19]. It was even shown in [27] on the MIRFLICKR data set [13] that while image categorization outperforms tag propagation when using properly labeled manual annotations, the latter outperforms when using the noisy Flickr tags as training labels.

In this paper therefore we are building on the ideas of TagProp [11], a nearest neighbor model which additionally allows for integrated metric learning. TagProp is a probabilistic method that predicts tags by taking a weighted combination of the tag absence/presence among neighbors. It generalizes the approach of [19], by learning a weight for each neighbor (based on its distance) by maximizing the likelihood of annotations in a set of training images. The main difference, with our paper, resides in the fact that in [11] the nearest neighbor images are gathered based only on visual similarities. On the contrary, we propose to use also the available textual information around these images (either the tags or the full captions) in order to improve the visual auto-annotation. Even though the new image contains only visual information, we are able to exploit the textual modality in the database to improve the performance. This is done with the integration of trans-media pseudo relevance feedback [5, 18] in the weighted nearest neighbor approach.

To evaluate our models and to compare to previous work, we use two data sets – Corel 5k and IAPR TC12. We compare our models with the results published in [11]. On both data sets we show that the new approach outperforms the original TagProp method.

The rest of the paper is as follows. In Section 2 we give some further background on image auto annotation and pseudo-relevance feedback methods which are closest to and inspired our method. In Section 3 we describe our proposed method in more detail and in Section 4 we give experimental evaluation and show excellent results. Finally, we conclude our paper in Section 5.

2 Related work

In this section we discuss models for image annotation and keyword based retrieval most relevant for our work. We identify four main groups of methods: those based on topic models, based on mixture models, discriminatively trained ones and nearest neighbor type models.

Topic based models use latent Dirichlet allocation, probabilistic latent semantic analysis, or hierarchical Dirichlet processes [2, 20, 29]. They model the annotated images as samples from a specific mixture of topics, where each topic is a distribution (most often Gaussian) over image features and annotation words (generally multinomial). Methods inspired by machine translation [7], where visual features are translated into the annotation vocabulary, can also be seen as topic models, where one topic is used per visual descriptor type. Although conceptually attractive, their expressive power is limited by the number of topics. Furthermore, these techniques function on image region level, and hence require as input labeled regions.

A second family of methods uses mixture models to define a joint distribution over image features and annotation tags. These models can be seen as non-parametric density estimators over the co-occurrence of images and annotations. To annotate a new image, these models compute the conditional probability over tags given the visual features by normalising the joint likelihood [3, 8, 14, 15]. As above, generally Gaussian mixtures are used to model visual features, while the distributions over annotations are multinomials or separate Bernoullis for each word.

Both families of generative models are critisized because maximizing the generative data likelihood might not be necessarily optimal for predictive performance. Therefore, alternatively, discriminative models for tag prediction were proposed in [6, 9, 12] that learn a separate classifier for each potential tag. This is equivalent to multi-class multi-label image categorization problem, and hence different learning methods can been used to train the classifiers, including support vector machines, Bayes point machines, etc.

Given the increasing amount of training data that is currently available, local learning techniques are becoming more attractive as a simple yet powerful alternative to parametric models. Examples of such techniques include methods based on label diffusion over a similarity graph of labeled and unlabeled images [17, 22], or learning discriminative models in neighborhoods of test images [30]. A simpler ad-hoc nearest neighbor tag transfer mechanism was recently introduced [19], showing state-of-the-art performance. There, nearest neighbors are determined by the average of several distances computed from different visual features. As a generalization of this method, Guillaumin *et al*. in [11] proposed TagProp that learns the weights for each neighbor (based on its distance) by maximizing the likelihood of annotations in a set of training images.

Reusing the information of nearest neighbors of a test images is the core of the pseudo-relevance feedback or query expansion mechanism that has been used widely both in text and image retrieval [4]. It was originally proposed in the context of text retrieval [25] and the main idea is to extend the initial query, with information taken from relevant documents. Since it is not *a-priori* known which documents are relevant to a query, pseudo-relevance feedback models use the top k retrieved documents with a predefined k and extracts information from them to enrich the query and do a more robust search.

The trans-media pseudo-relevance models [1, 5, 18] are extensions of these models, where the similarity functions used in the two retrieval steps are based on different modalities. For example, we start with a query image and select the k most similar images from the database. Then, the text associated with the top k images is used to re-rank the documents according to their textual similarity. These models have shown significant improvement on retrieval performance in multi-modal databases [1, 5].

In this paper, we propose to combine trans-media pseudo-relevance models with the TagProp auto-annotation method proposed in [11] in order to improve image auto-annotation performance.

3 Image Auto-annotation

The goal is to develop a method which is able to accurately predict the relevance of a concept for a given image based on the tags of the most similar images in a database. This process is known as tag propagation and is inspired by recent state-of-the art approaches [8, 11, 14, 19]. It is clear, that the quality of the set of images that is used to predict the labels of a new image is primordial. Therefore we propose to take advantage of the available textual data (in form of

the given annotations and captions) of the images and use the trans-media pseudo relevance feedback model as described in [1] to improve the quality of this relevant image set. This model is included in the weighted nearest neighbor approach proposed by Guillaumin *et al.* in [11] in order to simultaneously learn weighting parameters both for mono-modal and multi-modal distances.

In the next section we briefly describe the TagProp method [11], which outperform the methods of [8, 14, 19]. In Section 3.2 we present how we include the Trans-media pseudo relevance feedback model.

3.1 TagProp: Distance Based Nearest Neighbor Tag Prediction

To model image annotations, TagProp uses a Bernoulli model for each keyword, because keywords are either present or absent. Let $y_{it} \in \{-1,+1\}$ denote the absence/presence of tag t for image i, hence encoding the image annotations. The presence prediction $p(y_{it}=+1)$ for tag t from image i is defined as a weighted sum over the training images, indexed by j:

$$p(y_{it}=+1) = \sum_j p(y_{it}=+1|j)\; p(j|i), \quad \text{with} \quad p(y_{it}=+1|j) = \begin{cases} 1-\varepsilon & \text{if } y_{jt}=+1 \\ \varepsilon & \text{otherwise} \end{cases} \tag{1}$$

the ε is a technicality to avoid zero prediction probabilities, and in practice we set $\varepsilon = 10^{-5}$.

The probability to use image j as a neighbor for image i, $p(j|i)$ can be defined using image rank (i.e. image j is the k-th neighbor of image i) or image distance (i.e. using d_{ij} the distance between image i and image j). While the performance does not depend much on this choice [27], we prefer the distance based interpretation. This interpretation has the advantage that the weights depend smoothly on the distance, which allows for metric learning, and there is only a single parameter for each base distance.

$$p(j|i) = \frac{\exp(-\mathbf{w}^\top \mathbf{d}_{ij})}{\sum_{j' \in \mathbf{J_i}} \exp(-\mathbf{w}^\top \mathbf{d}_{ij'})}. \tag{2}$$

where $\mathbf{J_i}$ can be the whole data set or the subset of the J most similar images to i (all other weights are considered 0), $\mathbf{d}_{ij}$ is a vector of different base distances (visual in the original TagProp) between image i and j, and $\mathbf{w}$ controls the exponential decay.

To estimate the parameter vector $\mathbf{w}$, that controls the probability $p(j|i)$, the log-likelihood of the predictions of training annotations is maximized. Taking care to set the weight of training images to themselves to zero, i.e. $p(i|i)=0$, the objective is to maximize

$$L = \sum_i \sum_t c_{it} \ln p(y_{it}), \tag{3}$$

where c_{it} is a cost that takes into account the imbalance between keyword presence and absence:

$$c_{it} = \begin{cases} \frac{1}{n^+} & \text{if } y_{it}=+1 \\ \frac{1}{n^-} & \text{if } y_{it}=-1 \end{cases} \tag{4}$$

where n^+ and n^- are the total number of positive respectively negative labels. This cost weighting is used because in practice, there are many more tag absences than presences, and

absences are much noisier than presences. Indeed, most images are annotated with only a subset of all possible relevant keywords.

Notice that [11] in addition proposes an extended model which uses word-specific logistic discriminant models. While this extension is intended to boost the recall, in general the performance on mean average precision (**MAP**) are almost equal [27]. Therefore, in this paper we consider the former one as baseline and we compare our method to it. Nevertheless, the extension proposed here can easily be integrated with the word-specific models.

3.2 TagProp extended with Trans-Media Pseudo Relevance Feedback

In this section we present how we include the trans-media pseudo relevance feedback model into the weighted nearest neighbor model from the section above. The ideas for the trans-media pseudo-relevance model have shown excellent performance in multi-modal document retrieval [1, 5]. Before describing the integrated model we briefly remember the main principle. Trans-media pseudo-relevance feedback is an extension of the well-known relevance feedback principle, or query expansion, for multi-modal databases. The idea is that the first retrieval step is done in one modality (e.g. visual), and the second step is performed in another modality (e.g. textual), and this new modality is used to re-rank. Hence, the trans-modal distance d^{VT} between image i and j becomes

$$d^{VT}(i,j) = \sum_{k \in N_i^V} d^V(i,k)\, d^T(k,j), \tag{5}$$

where $d^T(k,j)$ is the textual similarity between the texts associated with image k and image j, and N_i^V is set of images retrieved using visual similarity (in the first) step with the query i. This is equivalent in our context to the $K = |N_i^V|$ nearest neighborhood of the image i based on visual distances.

We can see that Eq. 5 defines a new (cross-modal) distance between i and j that actually can replace the distance in Eq. 2 or it can be simply added to the linearly combined distances with an additional weight.

However, we go beyond this simple combination by generalizing the Eq. 5 as follows:

$$d_d^{VT}(i,j) = \sum_k \gamma_{dk}\, d_d^V(i,k)\, d^T(k,j), \tag{6}$$

where the subscript d means we used the d^{th} base distance $d_d^V(i,k)$ from the vector $\mathbf{d}_{ik}$, and γ_{dk} is used to weight the neighbors of the trans-media pseudo-relevance step. If we use $\gamma_{dk} = 1$ for all k, we obtain an equally weighted distances as in Eq. 5. The cross-media distance d^{VT} is used just as the other visual distances. To combine multiple cross-modal distances we define $f_{ij}^{VT} = \mathbf{w}_{VT}^\top \mathbf{d}_{ij}^{VT}$, where $\mathbf{w}^{VT}$ is the weighting vector, and $\mathbf{d}$ is a vector represention of the distances. Similarly, for the visual distances we define $f_{ij}^V = \mathbf{w}_V^\top \mathbf{d}_{ij}^V$. The probability that image j is a neighbor of image i based on both their visual and cross-modal similarities becomes:

$$p(j|i) = \frac{\exp(-(f_{ij}^V + f_{ij}^{VT}))}{\sum_{j' \in J_i} \exp(-(f_{ij'}^V + f_{ij'}^{VT}))}. \tag{7}$$

This can be replaced in the objective function to be maximized (Eq. 3). Whe refer to this method as Linear Transmedia Pseudo Relevance Feedback (**LTP**).

A drawback of this method could be that there are quite a lot of parameters to estimate, especially for large neighborhoods. We could introduce additional constraints to improve the generalization properties of the model. Such as a non-negativity constraint on the coefficients, or even an ordering constraint where we assume that the contribution of neighbor $i+1$ can not exceed that of neighbor i, i.e. $\gamma_i \geq \gamma_{i+1}$. Both these constraints define a convex feasible set.

Alternatively, we propose a second model which satisfies non-negativity and ordering constraints by construction, where we use the softmax function on $d_d^V(i,k)$ to define:

$$d_d^{VT}(i,j) = \sum_k \tilde{d}_d^V(i,k)\, d^T(k,j), \quad \text{where} \qquad \tilde{d}_d^V(i,k) = \frac{\exp(-\gamma_d\, d_d^V(i,k))}{\sum_{k'} \exp(-\gamma_d\, d_d^V(i,k))}. \tag{8}$$

This second model has the advantage that, for each visual distance d, it only has a single parameter γ_d opposed to K parameters for the first formulation. We refer to this method as Softmax Transmedia Pseudo Relevance Feedback (**STP**).

3.3 Learning the parameters of the model

For optimizing the parameters we directly maximize the log-likelihood using a projected gradient algorithm. The gradient descent procedure alternates over the steps described in Algorithm 1.

while *not converged* **do**
  if *STP* **then**
    minimize log-likelihood w.r.t. $\Gamma = \{\gamma_d\}$;
    compute $\mathbf{d}^{VT}$ given Γ;
  end
  minimize log-likelihood w.r.t. $\mathbf{w}_{VT}$ and $\mathbf{w}_V$, using $\mathbf{d}^{VT}$ and $\mathbf{d}^V$;
  calculate log-likelihood with $\mathbf{w}_{VT}, \mathbf{w}_V, \mathbf{d}^{VT}$ and $\mathbf{d}^V$;
  check for convergence;
end

Algorithm 1: *Pseudo Code for Iterative Learning Tagprop*

The derivatives of the model are described in Appendix A.

In the case we use the linear pseudo-relevance feedback (Eq. 6), we use each neighbor k from the first step as separate distance. Hence γ_{dk} is merged with w_d^{VT} to get a single weight parameter. This allows the direct use of the original TagProp method, however with a larger number of distances: $|d^V| \times (1+k)$.

4 Experiments

In this section we present a comparative evaluation of our two models (**LTP** and **STP**) with the original TagProp [11] on two publicly available data sets: Corel 5k [7] and the IAPR TC12 [10].

For a better comparison with the original TagProp we use the same visual features as in [11] that are available for download [1]. Just to summarize them briefly, they are 15 distinct

[1] http://lear.inrialpes.fr/data

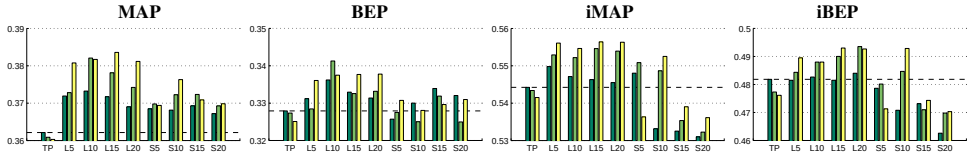

Figure 1: Linear and Softmax Pseudo Relevance Feedback Models using different K (indicated L5 (S5) for **LTP** (**STP**) with $K = 5$), and different sizes of neighborhood $J = \{200, 400, 1000\}$ indicated by the different bars, on Corel 5K dataset, using the tag distance as d^T.

descriptors: the Gist descriptor [21], 6 color histograms for RGB, LAB and HSV representations, and 8 bag-of-word histograms: 2 extraction methods x 2 descriptors x 2 layouts, where points are extracted using a dense grid or Harris-Laplacian interest points, the SIFT and robust hue descriptor [26] are used, and we use two spatial layouts, the whole image or three horizontal regions of the image. To compute the visual distances from the descriptors we follow [11, 19] and use L2 as the base metric for Gist, L1 for global color histograms, and χ^2 for the others. Besides these collection of 15 descriptors we often use them equally weighted and averaged and refer to that as JEC distance. This is the visual distance we used in most experiments and hence d^V is JEC and $d = 1$ whenever it is not precised otherwise.

In order to compare directly to TagProp we use the available annotations to define a textual distance. As textual features we use intersection over union of the set of tags of two images, $d^T_{kj} = 1 - |Y_k \cap Y_j|/|Y_k \cup Y_j|$, with $Y_k = \{t | y_{kt} = +1\}$, we refer to these distance as the tag distance. We also experimented with the classical TF/IDF (term frequency over inverted document frequency). Besides, we include experiments where we use the JEC distance as d^T, which results in visual pseudo-relevance feedback.

Finally, as the IAPR data set contains free text descriptions (we call them here captions), we used the language model proposed in [24] to represent these texts after pre-processing (tokenization, lemmatization and standard stop-word removal). The cross-entropy function was used as textual similarities between two image caption (see e.g [5] for details) and we refer to it as the text distance. This distance is used as d^T in the trans-media distance d^{VT}.

Image auto-annotation is usually evaluated measuring the keyword based retrieval of the system. To measure this performance we use the widespread mean average precision (**MAP**) and break-even point precision (**BEP**) over keywords. **MAP** is obtained by computing for each keyword the average of the precisions measured after each relevant image is retrieved. **BEP** (or R-precision) measures for each keyword (tag) t the precision among the top n_t relevant images, where n_t is the number of images annotated with this keyword in the ground truth. To evaluate the performance for annotating, we inverse these measures, and calculate **iMAP** and **iBEP**, where instead of calculating precision over ranked images and averaging over keywords, we calculate precision over ranked keywords, and average over all images.

4.1 Corel 5k dataset

In this section we perform experiments on the Corel 5K dataset. The dataset contains around 5000 images with manual annotation between 1 to 5 keywords. The images are annotated with for the purpose of keyword-based retrieval.

In Fig. 1 we show the performance of the original TagProp (TP) compared to our Trans-

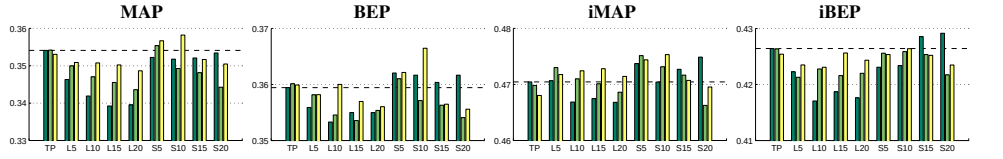

Figure 2: Linear and Softmax Pseudo Relevance Feedback Models using different K (indicated L5 (S5) for **LTP** (**STP**) with $K = 5$), and different sizes of neighborhood $J = \{200, 400, 1000\}$ indicated by the different bars, on IAPR TC-12 dataset.

Media Pseudo Relevance extensions, using the tag distance in d^{VT}. It shows that most parameter configurations and using any of the performance measures we significantly outperform the baseline TagProp. When comparing TagProp (using J=200, **MAP** 36.2) with our method (using **LTP**, K=15, J=1000, **MAP** 38.4%), on the AP per keyword (260 in total) we see that in 144/26/90 cases our method outperform/equals/underperforms TagProp. Furthermore, the figures show that **LTP** generally outperforms **STP** on this dataset. Finally, if we increase the neighborhood size J (indicated by different bars) the performances increases in our case while for TagProp slightly decreases.

In Table 1 we further compare the performance of the **LTP** and **STP** with different possibilities we can use for the distance d^T in Eq. 6. We can use again the visual distance (JEC), which makes it a visual pseudo-relevance feedback model, the tag distance, and finally combine the two. While using visual pseudo-relevance feedback performs similar to direct TagProp, the trans-media model clearly improves the retrieval and annotation performance. The combination of the two improves further on **iMAP**, while scoring equal or lower on the other measures. Just as in Fig. 1, **LTP** seems to outperform **STP** for these settings.

Table 1: Performance of different d^T distances, using $J = 1000$, and $K = 20$.

	LTP				**STP**			
	MAP	BEP	iMAP	iBEP	MAP	BEP	iMAP	iBEP
TagProp	36.0	32.5	54.2	47.6				
$d^T = \{\text{Jec}\}$	36.0	32.5	54.2	47.8	36.0	32.5	54.2	47.8
$d^T = \{\text{Tag}\}$	38.1	33.8	55.6	49.3	37.0	33.1	53.6	47.0
$d^T = \{\text{Jec, Tag}\}$	37.9	33.9	55.5	49.7	36.6	32.9	53.7	47.2

4.2 IAPR TC12 dataset

In this section we show experiments on the IAPR TC12 dataset. It contains about 20.000 images accompanied with descriptions, the annotation keywords are the common nouns of these descriptions extracted using natural language processing techniques.

In Fig. 2, we give an overview of the results of **LTP** and **STP** using the tag based textual features. We see that the TagProp baseline is much harder to beat in this setting than in case of the Corel 5K dataset. Also, on this dataset **LTP** model is clearly outperformed by the **STP** model.

For this dataset, we also have full captions available, which gives a different d^T measure. Although we note that the tags depend on the captions, therefore the tag and text distances are likely to be very similar. In Table 2 we show the performance when using the different pseudo-relevance distances, and combining them. We use $d^T = \{\text{JEC, Tag, Text}\}$, and the

Table 2: Combining different d^T distances, $J = 400$, and $K = 10$

	LTP				STP			
	MAP	BEP	iMAP	iBEP	MAP	BEP	iMAP	iBEP
TagProp	35.4	36.0	47.0	42.6				
$d^T = \{\text{Jec}\}$	35.1	36.0	46.7	42.2	35.1	36.0	46.7	42.3
$d^T = \{\text{Tag}\}$	34.7	35.5	47.1	42.3	35.6	36.3	47.4	42.7
$d^T = \{\text{Text}\}$	34.9	35.9	47.5	42.2	35.9	36.3	48.0	42.8
$d^T = \{\text{Tag, Text}\}$	34.7	35.8	47.2	42.1	35.7	36.5	47.9	43.0

Table 3: Combining 4 different base distances d^V, using $J = 400$, and $K = 10$

	LTP				STP			
	MAP	BEP	iMAP	iBEP	MAP	BEP	iMAP	iBEP
TagProp	35.7	36.1	49.0	44.1				
$d^T = \{\text{Jec}\}$	35.0	35.3	48.6	44.1	35.0	35.6	48.6	44.0
$d^T = \{\text{Tag}\}$	36.0	36.7	49.6	44.6	35.6	36.1	49.2	44.4
$d^T = \{\text{Text}\}$	36.4	36.7	49.6	44.3	35.7	35.7	49.5	44.2
$d^T = \{\text{Tag, Text}\}$	36.2	36.6	49.9	44.8	35.8	36.6	49.8	44.6

combination of Tag and Text distance. Using $d^T = \text{Text}$ we obtain the highest scores, improving around .5% the retrieval scores, and up to 1% the **iMAP**. Comparing on the AP per keyword (291 in total), between TagProp (using J=200) and our method (using text distance, **STP**, J=400,K=10), our method outperforms/equals/underperforms in 168/3/120 cases.

Finally, in Table 3 we show the performance when we use several base distances, so we are including the metric learning properties of TagProp. For this experiment we used the 4 distances from TagProp with the highest weights when learned with the 15 described distances (see [27] for an overview of the weights). The used features are the GIST, Dense-SIFT, Harris-SIFT, and Dense-SIFT-V3. Surprisingly, in this case the **LTP** model outperforms the **STP** model. Also we can see that using **LTP** with either the Tag or Text distance we improve on all performance measures.

5 Conclusion

We have introduced two models to use trans-media pseudo-relevance feedback for image auto-annotation. The two models (a linear and a softmax model) were integrated into TagProp, which is a probabilistic nearest-neighbor approach for image auto-annotation. Hence, we obtained an extended model which combines visual distances between two images with cross-modal visual-textual distances. The model further allows for metric learning, and the parameters can be trained in a discriminative manner using a log-likelihood optimization.

Our experiments show that we consistently outperform the state-of-the-art baseline of TagProp. On the Corel 5K dataset, we make a notable improvement in both keyword retrieval performance (measured by **MAP** and **BEP**) and in image annotation performance (measured by **iMAP** and **iBEP**). On the more challenging IAPR TC12 dataset, we also obtained slight improvements using a single visual distance. However, we have shown that when we further include metric learning by incorporating several visual distances both in the visual and in the cross-modal part, the improvements were up to 1% over our state-of-the-art baseline TagProp. To conclude, we have shown that using the available textual information around the images of the dataset (either the tags or the full captions) improves visual auto-annotation for both keyword based retrieval and annotation prediction

A Derivatives

In the case of **LTP** we can directly maximize the log-likelihood using a projected gradient algorithm. The gradient of the log-likelihood Eq. 3 with respect to the $\mathbf{w}$ equals:

$$\frac{\partial L}{\partial \mathbf{w}} = \sum_{i,j} C_i (p(j|i) - \rho_{ij}) \mathbf{d}_{ij}, \qquad \rho_{ij} = \sum_t \frac{c_{it}}{C_i} p(j|y_{it}), \tag{9}$$

where $C_i = \sum_t c_{it}$, and $\mathbf{w}$ is the collection of the $|d^V| \times (1+k)$ weighting parameters (w_d^V and $w_{dk}^{VT} = w_d^{VT} \gamma_{dk}^V$). To reduce the computational cost, we compute the pairwise distances over a large set of J neighbors, and assume the remaining ones to be zero. For each image, we include J neighbors such that we maximize the number of neighbors from each base distance (or base distance and text distance combination). In this way we are likely to include all images with large $p(j|i)$ regardless of the distance combination using $\mathbf{w}$ and γ that is learnt.

In the case of **STP**, to update γ_d we can use gradient descent with:

$$\frac{\partial L}{\partial \gamma_d} = \sum_{i,j} C_i (p(j|i) - \rho_{ij}) \frac{\partial f_{ij'}^{VT}}{\partial \gamma_d} \tag{10}$$

$$\frac{\partial f_{ij}^{VT}}{\partial \gamma_d} = w_d^{VT} \sum_k \tilde{d}_d^V(i,k)\, d_d^V(i,k) \left[d_d^{VT}(i,j) - d_d^T(k,j) \right], \tag{11}$$

where ρ_{ij} equals to the definition in Eq. 9.

References

[1] J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka, and J. Renders. XRCE's participation to ImageCLEF 2008. In *Working Notes of the 2008 CLEF Workshop*, 2008.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.

[3] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *PAMI*, 29(3):394–410, 2007.

[4] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

[5] S. Clinchant, J. Renders, and G. Csurka. XRCE's participation to ImageCLEFphoto 2007. In *Working Notes of the 2007 CLEF Workshop*, 2007.

[6] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using SVM. In *Proceedings Internet imaging (SPIE)*, volume 5304, 2004.

[7] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[8] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[9] D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *PAMI*, 30(8):1371–1384, 2008.

[10] M. Grubinger. *Analysis and evaluation of visual information systems performance*. PhD thesis, Victoria University, Melbourne, Australia, 2007.

[11] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[12] T. Hertz, A. Bar-Hillel, and D. Weinshall. Learning distance functions for image retrieval. In *CVPR*, 2004.

[13] M. Huiskes and M. Lew. The MIR Flickr retrieval evaluation. In *MIR*, 2008.

[14] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR*, 2003.

[15] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[16] X. Li, L. Chen, L. Zhang, F. Lin, and W. Ma. Image annotation by large-scale content based image retrieval. In *ACM Int. Conf. on Multimedia*, 2006.

[17] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *PR*, 42 (2):218–228, 2009.

[18] N. Maillot, J.-.P. Chevallet, V. Valea, and J. H. Lim. IPAL inter-media pseudo-relevance feedback approach to ImageCLEF 2006 photo retrieval. In *Working Notes of the 2006 CLEF Workshop*, 2006.

[19] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[20] F. Monay and D. Gatica-Perez. PLSA-based image auto-annotation: Constraining the latent space. In *ACM Multimedia*, 2004.

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[22] J. Pan, H. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *SIGKDD*, 2004.

[23] F. Perronnin, J. Sanchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *CVPR*, 2010.

[24] J. Ponte and W. Croft. A language modelling approach to information retrieval. In *ACM SIGIR*, pages 275–281, 1998.

[25] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[26] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006.

[27] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with TagProp on the MIRFLICKR set. *ACM Multimedia Information Retrieval*, 2010.

[28] X. Wang, L. Zhang, F. Jing, and W.-Y. Ma. AnnoSearch: Image auto-annotation by search. In *CVPR*, 2006.

[29] O. Yakhnenko and V. Honavar. Annotating images and image objects using a hierarchical Dirichlet process model. In *Workshop on Multimedia Data Mining ACM SIGKDD*, 2008.

[30] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.